Subject analysis: the critical first stage in indexing

By: Clara M. Chu and Ann O'Brien

Chu, Clara and O'Brien, Ann. "Subject Analysis: The Critical First Stage in Indexing," *Journal of Information Science*, 19(6): 439-54, 1993.

Abstract:
Studies of indexing neglect the first stage of the process, that is, subject analysis. In this study, novice indexers were asked to analyse three short, popular journal articles; to express the general subject as well as the primary and secondary topics in natural language statements; to state what influenced the analysis and to comment on the ease or difficulty of this process. The factors which influenced the process were: the subject discipline concerned, factual tiss. subjective nature of the text, complexity of the subject, clarity of text, possible support offered by bibliographic apparatus such as title, etc. The findings showed that with the social science and science texts, the general subject could be determined with ease, while this was more difficult with the humanities text. Clear evidence emerged of the importance of bibliographical apparatus in defining the general subject. There was varying difficulty in determining the primary and secondary topics.

Article:
1. Introduction
Most studies about indexing are in fact about indexes. This sentiment was expressed by Jones in 1976 and still holds true [1]. Svenonius characterises studies on indexing into (a) those that contribute to decision making as evaluative research (concerned with the performance of indexing systems) and (b) those that are concerned with problem solving as developmental research (primarily concerned with the designing of indexing systems) [2]. With the wide range of indexing systems available and their importance for retrieval, it is hardly surprising that a detailed examination of their quality, efficiency and effectiveness is of central concern.

However, the result of this focus means that one part of the indexing process, i.e., subject analysis, has been neglected to the extent that insufficient emphasis has been given to a close examination of how text is analysed. This process underlies all research whether evaluative or developmental. It appears that there is an assumption that we have learnt all we need to know about the analysis process when applied to text and that the process is problem-free (i.e., the analysis is without difficulties, so that co-extensiveness with, and consistency of, the subject are achieved between and across indexers). Work has been done outside the area of library and information studies but although valuable, much of this has concentrated on comprehension of text and parts of text and lacks the contextual focus which indexing requires [3]. With the current surge of interest in image retrieval, the case of text risks being ignored, when it is clear that many problems still persist.
Vickery gives a useful definition of subject analysis which is adopted in this study:

> "Analysis of information' here means deriving from a document a set of words that serves as a condensed representation of it. This representation may be used to identify the document, to provide access points m literature search or indicate its content, or as a substitute for the document" [4]

However, it is true to say that in practice, the process of translating the subject content of text into index entries (of whatever system) is more often than not seen as a single process. It is the contention of the authors of this paper that such an approach misses out on an essential stage in the process which takes place at the subject analysis phase. Subject analysis of the text occurs first and the topics decided upon are then translated into the relevant indexing system.

Stages of the indexing process:

1. subject analysis of text,
2. expression of the subject content in the indexers' words,
3. translation into an indexing vocabulary,
4. expression of the subject in index terms.

For experienced subject cataloguers and indexers, the first three steps may happen almost simultaneously but this is not to imply that they are a single activity. Indeed, if indexers approach a text wholly with the intention of fitting the subject matter into their system vocabulary, they may lose some of the nuances which could enhance the subsequent index terms.

This study attempts to understand the process of subject analysis at the macro- and micro-levels: what decisions are made, what guides these decisions and what helps or hinders the process (as at Stages 1 and 2 above) without the constraints of fitting them into any particular system. At the macro-level, the relationship between text and subject has rarely been examined by itself in a way that attempts to capture a general natural language statement from the indexer. A natural language statement may have a tendency to aim towards co-extensiveness with the subject or at least attempt, as Vickery says,, to give a "condensed representation" of what it is about. If the indexer is not immediately concerned with trying to fit the response into a structured vocabulary, the analysis is more likely to reveal the true nature of how people analyse text. If pushed further, the concepts expressed might, if appropriate, be ranked as primary, secondary, etc. In other words, the process could be examined at the micro-level giving even more detailed insight into the essential elements of the first stages of the indexing process. By allowing a natural language approach to both levels, questions of emphasis, balance and specificity emerge in an unencumbered way.

The process of analysis encompasses various activities, i.e., scanning the text, identifying the subject concepts, organising these into possible priorities. These happen within the context of several inter-related influencing factors: personal, procedural, document-related and environmental [5]. As this study was not based on any particular indexing system but concentrated on the subject analysis phase only, the factors to be studied excluded the procedural, which has to do with the constraints of the indexing system used. Environmental factors (e.g., physical surroundings, work quotas, space, noise) were of minimal concern, as both groups of participants were in a partially- controlled environment. The tasks were performed in groups in a classroom. Personal factors, such as age, gender, professional experience, etc. and document-related factors, such as subject matter, layout, style, etc. were the basis of this research and are discussed in detail below.

At some stage in the process, it is necessary to say what the text is about. There are several approaches to this large question of determining aboutness in text [6,7]. The theoretical approach offers behavioural and/or linguistic interpretation [8] and the practical approach is to be found in textbooks or manuals on indexing and offers historical background and guidance in indexing with an emphasis on individual systems and using them to assign terms [5]. For the purposes of this study, analysis was done by novice indexers who were encouraged to give their own generalised statements of the subject content and to say how they arrived at these. Their views were compared with the authors' analyses, which were taken to be definitive and as such, a practical approach to aboutness predominates.

Another concern in the indexing process as a whole is that of consistency. There are extensive studies which show inconsistency on the part of indexers in their choice of indexing system terms for similar text. Choice of the appropriate terms is heavily dependent on the result of the subject analysis stage and should not be wholly attributed, as in many indexing consistency studies, to indexer experience and particulars of indexing systems. Consistency is critical at all stages of the indexing process. Inconsistency during subject analysis, the focus of this study, can be the result of expressing subject concepts in different ways and at different levels of specificity which can lead to varying terms chosen for similar documents [9].

Difficulties with aboutness and inconsistency result in poor quality indexing, which in turn hinders the quality of information retrieval. Even more reason then to address this basic but critical issue: to try to understand, in a practical context, what influencing factors predominate in subject analysis; how they might be measured; and if, without the constraints of any particular indexing system, subject analysis results in the articulation of concepts which are generic to the indexing process. To this purpose two major questions were addressed. The first concerned the performance of the task of analysing the subject of text, the second concerned the ease with which the task was accomplished. For both of these questions, the factors which contributed to or hindered the outcome were examined:

1 (a) Can novice indexers successfully perform subject analysis (i.e., derive the subject of texts which match with definitive subjects determined by the authors of this paper)? In this study, subject analysis involves the identification of the central concept (i.e., the general subject) and the prioritising of the subject into primary and secondary topics if appropriate.
(b) What helped novice indexers determine the subject of an article?
2 (a) What was the relative ease of performing subject analysis?
(b) What affected the ease or difficulty of the subject analysis experience?

Alongside these questions, a number of detailed issues concerning the influencing factors were addressed:
1. What role do document-related factors such as bibliographical apparatus and presentational layout play in determining the subject content? How much and what parts of the text are needed to determine its subject content?
2. It is generally accepted that it can be more difficult to determine the subject of text in the humanities than in the social sciences and sciences. Some of this difficulty may be related to the subjective t's. factual nature of the respective texts. The choice of texts in this study was influenced by this.
3. Also of concern were issues of emphasis and balance. Does the text have more than one central concept (i.e., is it a compound subject) and if so, is it possible to state which is the dominant one? Texts were chosen in which the subject matter could be expressed as primary and secondary topics.

Attention to these questions and issues is long overdue. It is expected that the findings will not only provide a better understanding of the subject analysis process but will harness interest in this basic indexing activity and raise further issues for investigation.

2. Methodology

Data were collected from novice indexers using a questionnaire. The word indexer is being used here to refer to anyone (e.g., cataloguer, indexer, classifier) who would be doing subject work (i.e., analysing the subject of an item, describing the subject in his/ her terms, and translating it into a specific system's language). Novice indexers were selected because they would have received some introduction to the task of subject analysis, but would not have had sufficient experience to bias their performance. The novice indexers in this study were library and information science students from the Graduate School of Library and Information Science (GSLIS) at the University of California, Los Angeles (UCLA) and the Department of Information and Library Studies (DILS) at Loughborough University of Technology (LUT). GSLIS students were taking the required master's course on subject analysis during Winter 1991 or 1992, and DILS students were taking the required second-year undergraduate course on subject analysis and indexing during Autumn 1991. The students were invited to participate in the study which required less than an hour of their time. The questionnaire was administered at the beginning of the course. One hundred and four students (69 from GSLIS and 35 from DILS) participated in the study.

The questionnaire administered to the participants required them to read three magazine articles, and to respond to questions regarding the subject of each article and the ease or difficulty experienced in analysing the subject. Short magazine articles were selected to keep the reading to a minimum and to enable participants interested in reading the full text to manageit within a short time. Other criteria used in the selection of articles were variability in subject discipline, emphases on different facets (e.g., personality, matter, energy,energy, space,

time), variability in clarity of subject, material which would be understandable without expertise in the subject area, and variability in the physical layout and bibliographic apparatus.

The first article was about economic expansion of South Africa into Black African countries, the second was on discoveries of Venus by the Magellan spacecraft, and the third was about personal experiences in searching for social and cultural identification [10-12]. The general subject of the articles as well as their primary and secondary topics were determined by the authors who have considerable professional, teaching and research experience in indexing. These subject statements

Table 1
Sex of participants

School	Sex		Total
	Female	Male	
GSLIS	53	16	69
DILS	27	8	35
Total	80 (77%)	24 (23%)	104

Table 2
Age of participants

School	Age (yrs)				Total
	< 25	25–34	35–44	45–54	
GSLIS	16	31	16	6	69
DILS	26	4	3	2	35
Total	42 (40%)	35 (34%)	19 (18%)	8 (8%)	104

participants with one year of LIS studies were DILS students who were scheduled to take their subject analysis course in the autumn term of their second year of a three-year undergraduate degree programme. The 6 participants with more than one year of LIS studies were part-time or mature GSLIS or DILS students.

For the most part the participants' undergraduate majors were evenly distributed across library and information studies, humanities (including

Table 3
Participants' years of LIS experience

Years of LIS experience	Total
No experience	23
1 year or less	26
> 1 year to 2.9 years	19
3 to 4.9 years	18
5 or more years	17
No response	1

Average years of LIS experience = 288.2/103 = 2.8.

Table 4
Participants' type of LIS experience

Type of LIS experience	Total
No experience	19
General library asst. duties	48
Technical assistant	16
Reference assistant	8
Other	7
No response	6

were deemed authoritative and became the standards for the study. The participants' responses were compared to these definitive statements.

The questionnaire contained closed-ended questions, questions with Likert scales, and open-ended questions. For each open-ended question the authors jointly developed categories and codes to represent responses. The authors jointly coded all open-ended responses.

3. Results

Sixty-nine master's students participated from GSLIS and thirty-five undergraduate students participated from DILS (n = 104). The sample represents a response rate of 85%. Women made up 77% of the participants from each school (see Table 1). Seventy-four percent of the participants were below 35 years of age (see Table 2). The GSLIS students tended to be a bit older (45% were in the 25-34 years category) than the DILS students (74% were in the below 25 years category).

Approximately half of the students had very little or no experience doing library and information (LIS) work (48% had one year or less experience) (see Table 3). The longest any one of the participants had worked in the field was 20 years. The average number of years of LIS experience the participants had was 2.8 (288.2/ 103). Regardless of the extent of experience, the type of LIS work was still limited to clerical and library assistant duties (see Table 4). Many of the subjects had general library assistant experience while only a few had technical services or reference experience. The seven participants with other experience held various positions that included management and account representative, shelver and children's programme designer, programmer and analyst, manager of branch or small library, translator and online searcher, and special services development officer.

The participants were students taking a required course in subject analysis; therefore, they were still in the early stages of their degree programmes. The 66 participants with one term of LIS studies were GSLIS students who were scheduled to take their subject analysis course in the second term of the first year in a two-year master's degree programme (see Table 5). The 32

Table 5
Participants' length of time in LIS studies

Years of LIS studies	Total
One term	66
One year	32
More than one year	6

Table 7
Article 1: General subject and factors helpful in determining the general subject

Factor							
General subject	Layout	Body	Layout and body	Pers. knowl.	Other	No resp.	Total
1	43	7	20	0	3	0	73
2	3	1	0	1	0	0	5
3	15	0	6	0	0	1	22
4	3	1	0	0	0	0	4
Total	64	9	26	1	3	1	104

Legend for "General subject":
1. Economic expansion of South Africa into Black African countries (Authors' choice).
2. Black African countries' economic expansion into South Africa.
3. Economic expansion of South Africa into Black African countries and *vice versa*.
4. Unrelated responses.

history and liberal arts), and social sciences (see Table 6). All the DILS participants were doing an undergraduate degree in information and library studies and 4 of the GSLIS master's students had an undergraduate degree in information and/ or library studies. Science and technology undergraduate majors had limited representation in the study sample.

3.1. Article one: Rake, Alan, South Africa Wants to Make Friends, New African (January 1991) 27
The first article dealt with a social science issue. After a reading of the article the authors described its subject as the economic expansion of South Africa into Black African countries. The response given by most of the participants (70%) when asked to describe the subject of article 1 in general terms, matched with the authors' (see Table 7). The second most cited response (21%) was that the article was about the economic expansion of South Africa into Black African countries and vice versa. From the overwhelming number who identified the general subject and from the limited variability of responses, the subject appears to be readily identifiable. In the case of the responses which did not match the authors', a broader subject was reported.

The participants overwhelmingly (64/ 104 = 62%) indicated that the layout helped them to determine the subject of the article (see Table 7). In this study the layout refers to the bibliographical apparatus (e.g., title, abstract, first and last paragraphs, keywords, illustrations, etc.) and the physical presentation of a work. There was an average of 2.3 layout elements reported per respondent (234/ 103) (see Table 8). The ones cited by 25 or more respondents included initial paragraphs (25), title (31), paragraph headings (32), and abstract (68). Both the body of the text and the layout were cited (26/ 104 = 25%) (see Table 7) as second most helpful. The body of the text emphasizes the intellectual rather than the physical content of a work. Responses in the other category included I really can't say and Nothing particularly.

The primary topic of article 1 was determined by the authors to be economic relations in Africa. The term economic relations included economic growth, expansion, opportunity, cooperation and the like. Again, an overwhelming majority (73/ 104 = 70%) matched the primary topic as determined by the authors (see Table 9). In contrast to what helped the participants determine the general subject of the article, 70% (72/ 103) indicated that the body of the text helped them to determine the primary topic of article 1. Seventeen percent of the participants (17/ 103) indicated that both the body and the layout helped them determine the primary topic and 14% (14/ 103) found the layout to be a helpful guide.

Table 6
Participants' undergraduate major

Undergraduate major	Total
Library and information studies	39
Humanities (incl. history and liberal arts)	32
Social sciences	29
Applied sciences	3
Other	1

The authors considered the secondary topic of article 1 to be *apartheid*. The most cited topic (40/104 = 38%) matched the authors' and the second most cited response (20/104 = 19%) was that there was no secondary topic (see Table 10). In a case where the primary topic was quite clear, such as with article 1, there was more variability in the identification of a secondary topic and only a 38% match with the secondary topic determined by the authors. Responses reported by 10 or more participants included *political change* (15), *sanctions* (11), and *economic agreement* (10). Again, in contrast to what helped the participants

Table 8
Articles 1–3: Detailed factors helpful in determining the general subject *

Factor	Article		
	1	2	3
Nothing	0	1	0
Running head	6	n/a	n/a
Title	31	32	26
Subtitle	n/a	23	n/a **
Abstract	68	n/a **	29
Paragraph headings	32	n/a	n/a
Extracts	0	1	4
Initial paragraphs	25	25	10
Structure	4	0	1
Selected parts of text	3	0	1
Body of text	36	45	47
Personal knowledge	2	2	0
First sentences of paragraphs	4	4	2
First paragraph of each section	2	n/a	n/a
Last paragraph	6	5	5
Last sentences of paragraphs	0	0	1
Keywords	11	6	6
Size of type	1	0	0
Illustrations	n/a	19	n/a
Illustration captions	n/a	10	n/a
Writing style	0	0	2
Other	3	5	2
No response	1	4	18
Ave. no. of factors assigned by respondents	234/103 = 2.3	179/100 = 1.8	137/86 1.6

* The data presented in this table represent "all" the factors the participants cited as helpful in determining the general subject of articles 1, 2 and 3. The responses corresponding to a participant for articles 1, 2 and 3 have been recoded to represent one factor for each respondent and are presented in Tables 7, 13 and 18, respectively.

** This element was not present in the article but was cited by one respondent.

Table 9
Article 1: Primary topic and factors helpful in determining it

Factor / Primary topic	Layout	Body	Layout and body	No resp.	Total
1	10	50	12	1	73
2	0	1	0	0	1
3	0	3	0	0	3
4	1	1	1	0	3
5	1	2	0	0	3
6	0	3	1	0	4
7	0	1	0	0	1
8	0	1	0	0	1
9	0	1	0	0	1
10	0	1	0	0	1
11	1	0	0	0	1
12	0	2	1	0	3
13	0	1	0	0	1
14	0	2	1	0	3
15	0	0	1	0	1
16	0	2	0	0	2
17	1	1	0	0	2
Total	14	72	17	1	104

Legend for "Primary topic":
1. Economic relations in Africa (authors' choice).
2. Profit.
3. Economics.
4. Economic cooperation of African countries.
5. Economic sanctions.
6. Economic friendship.
7. South African businessmen.
8. Economic invasion.
9. Financial relations.
10. Welcoming overtures of economic cooperation.
11. "Political correctness".
12. Trade.
13. Persuading trade partners.
14. Relations of South Africa.
15. Government relations.
16. Post-apartheid South Africa.
17. Collapse of apartheid.

determine the general subject of the article but concurrent with what helped them select the primary topic, 60% (59/99) indicated that the body of the text helped them to determine the secondary topic of article 1. Other elements which helped participants determine the secondary topic included the layout (4), both the layout and body of the text (14), personal knowledge (1), and subjective elements, such as, instinct (1).

When the participants were asked how important some aspects of the subject were in relation to each other, the responses reflected the extent of agreement the participants had in identifying

Table 10
Article 1: Secondary topic and factors helpful in determining it

Factor / Secondary topic	N/A	Layout	Body	Layout and body	Pers. knowl.	Subjective	No. resp.	Total
0	20	0	0	0	0	0	0	20
1	0	1	27	5	1	1	5	40
2	0	0	9	2	0	0	0	11
3	0	3	4	3	0	0	0	10
4	0	0	11	4	0	0	0	15
5	0	0	3	0	0	0	0	3
6	0	0	1	0	0	0	0	1
7	0	0	2	0	0	0	0	2
8	0	0	1	0	0	0	0	1
88	0	0	1	0	0	0	0	1
Total	20	4	59	14	1	1	5	104

Legend for "Secondary topic":
0, No secondary topic.
1, Apartheid (authors' choice).
2, Sanctions.
3, Economic agreement.
4, Political change.
5, "Political correctness".
6, Improved relations.
7, African countries and/or individual countries.
8, Economic influence of South Africa.
88, Other.

the primary and secondary topics and, accordingly, the extent of variability in responses (see Table 11). Seventy-eight percent (80/ 102) of the participants considered that the primary topic was "more important" or "a bit more important" than the other aspects of the subject (i.e., the primary topic clearly stood out as more important ; therefore, making it easy to identify). Eighty percent of the respondents (71/89) felt that the secondary topic was "as important" or "a bit less important" than the primary topic (i.e., the secondary topic was not considerably less important, therefore, making it difficult to identify). Seventy- two percent of the respondents (59/82) found that other aspects of the subject were "a bit more important" and "as important" as the secondary topic (i.e., the relative importance of other topics apart from the primary one made it difficult to select one as secondary). The last two findings reveal why there was variability in the responses for a secondary topic for article 1.

Finally for article 1, when the participants were asked how difficult it was to determine its subject, 64% (65/ 101) indicated that it was "somewhat easy" to "very easy" (see Table 12). The most frequent reasons they gave for the relative ease in identifying the subject were the layout (33/65 = 51%) and the clarity of the narrative (16/65 == 25%).). This finding parallels the finding for what helped guide participants in determining the general subject of article 1.

3.2. Article two: Lemonick, Michael D, A Restless Venus Unveiled, Time (October 8, 1990) 69

The second article dealt with a subject from the sciences. The authors described its subject as

Table 11
Article 1: Relative importance of different aspects of the subject

Topics / Rating	Primary vs. others	Secondary vs. others	Secondary vs. others
Less important	0	10	4
A bit less important	2	30	9
As important	20	41	30
A bit more important	44	6	29
More important	36	2	10
No response	2	15	22

Table 12
Article 1: Ease in determining the general subject and reasons for the degree of ease

Reasons / Ease	1	2	3	4	5	6	7	8	NR	Total
Very easy	0	0	0	7	1	10	0	0	4	22
Somewhat easy	0	0	2	26	0	6	0	0	9	43
Fair	0	0	0	2	0	3	1	3	8	17
Somewhat difficult	2	2	1	0	0	1	1	5	4	16
Very difficult	0	0	0	0	0	0	0	2	1	3
No response	0	0	0	0	0	0	0	0	3	3
Total	2	2	3	35	1	20	2	10	29	104

Legend for "Reasons":
1, Unfamiliarity with task.
2, Lack of familiarity with subject.
3, Familiarity with subject.
4, Layout.
5, Familiarity with subject and layout.
6, Clarity of narrative.
7, Poor physical presentation (e.g., bad reproduction).
8, Lack of clarity of narrative.
NR, No response.

the discoveries of Venus by the Magellan spacecraft. The response given by most of the participants (74/ 102 = 73%) when asked to describe its subject in general terms matched with the authors' (see Table 13). The second most cited response (25/ 102 = 25%) was geology of Venus. Parallel to the first article, the fact that an overwhelming number of respondents identified the general subject and there was a limited variability of responses indicated that the subject was readily identifiable.

Approximately half of the respondents (49/ 100 = 49%) indicated that the layout helped them to determine the general subject of the article (see Table 13). There was an average of 1.8 layout elements reported per respondent (179/ 100) (see Table 8). The ones cited by 23 or more respondents included subtitle (23), initial paragraphs (25), and title (32). Unique to this article is the inclusion of an illustration but interestingly enough only 19 respondents cited that it helped them in determining the article's general subject.

Table 13
Article 2: General subject and factors helpful in determining the general subject

Factor / General subject	N/A	Layout	Body	Layout and body	Pers. knowl.	Other	No resp.	Total
1	1	37	18	12	1	4	1	74
2	0	10	9	5	0	0	1	25
3	0	1	0	0	0	1	0	2
4	0	1	0	0	0	0	0	1
9	0	0	0	0	0	0	2	2
Total	1	49	27	17	1	5	4	104

Legend for "General subject":
1, Discoveries of Venus by Magellan spacecraft (authors' choice).
2, Geology of Venus.
3, Magellan spacecraft.
4, Unrelated response.
9, No response.

The body of the text was cited as second most helpful (27/ 100 = 27%) (see Table 13) and the combination of both the layout and body of the text was cited as third most helpful (17/ 100 == 17%). These findings on what helped respondents determine the subject of article 2 correspond with the findings for article 1 except that there is a bit less reliance on the layout and a bit more on the body as well as both the layout and body.

The authors identified the primary topic of article 2 as the discoveries of Venus by the Magellan spacecraft (i.e., the primary topic is the same as the general subject). Two topics were most frequently reported: the discoveries of Venus by the Magellan spacecraft (29/ 102 = 28%) and Venus (general, discoveries, exploration) (35/ 102 = 34%) (see Table 14). Unlike the first article, the primary topic noted by the participants did not overwhelmingly match the authors'. The difficulty in this case was determining whether the primary topic was one aspect of the general subject (in this case, the general subject was a compound subject) or the general subject itself.

Again, in contrast to what helped the participants determine the general subject of the article, the majority of the respondents (58/99 = 59%) indicated that the body of the text helped them to determine the primary topic of article 2 (see

Table 14
Article 2: Primary topic and factors helpful in determining it

Factor / General subject	Layout	Body	Layout and body	No resp.	Total
1	9	16	3	1	29
2	8	20	6	1	35
3	2	6	0	0	8
4	2	12	4	0	18
5	4	4	2	1	11
6	1	0	0	0	1
9	0	0	0	2	2
Total	26	58	15	5	104

Legend for "Primary topic":
1, Discoveries of Venus by Magellan spacecraft (authors' choice).
2, Venus (general, discoveries, exploration).
3, Geology of Venus.
4, Geographic and atmospheric aspects of Venus.
5, Magellan spacecraft.
6, Mapping of Venus.
9, No response

Table 15
Article 2: Secondary topic and factors helpful in determin it

Factor / Secondary subject	N/A	Lay-out	Body	Layout & body	No resp.	To-tal
0	1	0	0	0	18	19
1	0	2	7	4	2	15
2	0	1	3	0	0	4
3	0	11	23	11	2	47
4	0	2	1	0	0	3
5	0	4	6	3	1	14
9	1	0	0	0	1	2
Total	2	20	40	18	24	104

Legend for "Secondary topic":
0, No secondary topic.
1, Geology of Venus (authors' choice).
2, Discoveries of Venus by Magellan spacecraft.
3, Magellan spacecraft.
4, Venus research and scholarship.
5, Comparison of Venus to Earth.
9, No response.

Table 14). Twenty-six percent (26/99) found the layout helped them determine the primary topic and fifteen percent (15/99) indicated that both the body and the layout were helpful elements. These findings correspond to the same activity for article 1 except that in this case fewer relied on the body of the text, and a few more relied on the layout as well as both the body and the layout.

The authors considered the geology of Venus as the secondary topic of article 2. Approximately half of the participants (47/ 102 = 46%)) described the Magellan spacecraft as the secondary topic (see Table 15). Only 15% (15/ 102) stated the same secondary topic as the authors and 19% (19/102) reported that there was no secondary topic. In this case where the primary topic was not quite clear, there was variability in the identification of a secondary topic. For the most part, the respondents who reported the geology of Venus as the primary topic noted the Magellan spacecraft as the secondary topic, and vice versa. Again, as in previous cases where participants were asked what helped them select the primary and secondary topics, many (40/ 80 = 50%) indicated that the body of the text helped them to determine the secondary topic of article 2. Other elements which helped respondents determine the secondary topic included the layout (20/80

Table 16
Article 2: Relative importance of different aspects of tl subject

Topics / Rating	Primary vs. others	Secondary vs. primary	Secondary vs. others
Less important	0	10	1
A bit less important	3	33	13
As important	22	35	29
A bit more important	36	2	25
More important	41	4	11
No response	2	20	25

= 25%), and both the layout and body of the text (18/80 = 23%).

As found for article 1, when the participants were asked how important some aspects of the subject were in relation to each other, the responses for article 2 reflected the extent of agreement the participants had in identifying the primary and secondary topics and, accordingly, the extent of variability in responses (see Table 16). Ninety-seven percent (99/ 102) of the participants considered that the primary topic was "more important" (40%), "a bit more important" (35%), or "as important" (22%) than the other aspects of the subject (i.e., the primary topic clearly stood out as more important for most people; therefore, making it easy to identify, but in some cases the participants had to make a decision between comparable topics of importance).

Eighty-one percent of the respondents (68/ 84) felt that the secondary topic was "as important" or "a bit less important" than the primary topic (i.e., the secondary topic was not considerably less important; therefore, making it difficult to determine). Eighty-five percent of the respondents (67/79) found that other aspects of the subject were "a bit more important," "as important" or "a bit less important" than the secondary topic (i.e., the relative importance of other topics apart from the primary one made it difficult to select one as secondary). Compound subjects, as found in article 2, can create difficulties if indexers are required to separate them into simpler subjects of distinct importance.

When the participants were asked how difficult it was to determine the general subject of article 2, 90% (86/96) of the respondents indicated that it was "fair" to "very easy" (see Table 17). The most frequent reasons reported for the relative ease (i.e., "somewhat easy" and "very easy" responses) in identifying the subject were the clarity of the narrative (32/ 65 = 49%) and the layout (13/ 65 = 20%). This finding parallels the finding for what helped guide participants in determining the general subject of article 2 but it differs from the reasons for the relative ease in

Table 17
Article 2: Ease in determining the general subject and reasons for the degree of ease

Reason / Ease	0	1	2	3	4	6	7	8	NR	Total
Very easy	0	0	0	3	7	19	0	0	4	33
Somewhat easy	1	0	1	0	3	10	0	1	9	25
Fair	0	4	2	1	3	3	1	6	8	28
Somewhat difficult	0	0	0	0	0	0	0	4	4	8
Very difficult	0	0	0	0	0	0	0	2	0	2
No response	0	0	0	0	0	0	0	0	8	8
Total	1	4	3	4	13	32	1	13	33	104

Legend for "Reasons":
0, Interest in subject.
1, Unfamiliarity with task.
2, Lack of familiarity with subject.
3, Familiarity with subject.
4, Layout.
6, Clarity of narrative.
7, Poor physical presentation (e.g., bad reproduction).
8, Lack of clarity of narrative.

Table 18
Article 3: General subject and factors helpful in determining the general subject

Factor / General subject	Layout	Body	Layout and body	Style	Other	No resp.	Total
1	17	9	3	2	1	3	35
2	3	4	2	0	0	0	9
3	1	0	0	0	0	2	3
4	8	11	7	0	0	0	26
5	4	2	4	0	1	1	12
6	2	3	2	0	0	0	7
9	0	0	0	0	0	12	12
Total	35	29	18	2	2	18	104

Legend for "General subject":
1, Personal experiences in searching for social and cultural identification (authors' choice).
2, Effects of parenting by parents embracing the social revolution of the 1960s.
3, Generation gap.
4, Current self-perception of young people.
5, Life choices.
6, Twentysomething generation group.
9, No response.

identifying the subject of article 1. In the first case, there was more reliance on the layout than the clarity of the narrative.

3.3. Article three: Berger, Arion, Too Many Options ? A Twentysomething Woman Ponders her Future, Utne Reader (January/ February 1991) 60-2. Excerpted from L.A. Weekly (March 30, 1990)

The last article was a prose piece providing a personal perspective and can be considered to be a humanities document. The authors described its subject as personal experiences in searching for social and cultural identification. The most frequently cited response (35/92 == 38%) matched with the authors' (see Table 18) and the second most cited response (26/ 92 = 28%) reported a more generalised subject than that identified by the authors: current self-perception of young people. This link to the general subject determined by the authors shows that the respondents were closely connected in their analysis and determination of the general subject.

The respondents emphasised three sources which helped them to determine the general subject of the article. Forty-one percent (35/86) found the layout helpful, 34% (29/86) relied on the body of the text to help them determine the general subject, and 21 % (18/86) 86) relied on both the body and the layout. In line with the previous articles, there was more reliance on the layout but it was not prominent. There was an average of 1.6 layout elements reported per respondent ((137/86) 86) (see Table 18). The ones cited by 25 or more respondents included title (26) and abstract (29).

The authors determined the primary topic of article 3 to be personal experiences in searching for social and cultural identification (again, the secondary topic is the same as the general subject). The three most reported responses for primary topic were generalised related topics of the primary topic determined by the authors (see Table 19). They were current self-perception of young people (26/ 87 == 30%), twentysomething generation group (17/87 87 == 20%), and life choices (13/ 87 = 15%). The number of responses that matched the authors' was very low (8/87 = 9%).). Similar to the situation with article two, the general subject was a compound one but unlike that article, it was not a matter of selecting from mainly two aspects of the subject. The subjective and personal nature of this article generated primary topics which were broader than that determined by the authors.

The body of the text continues to be the most helpful source (48/ 75 = 64%)) in guiding the respondents to determine the primary topic of an article (see Table 19). Helpful sources which were cited less frequently were both body and layout (12/75 = 16%) and layout (11/ 75 = 15l7c).

The authors considered the effects of' patenting by parents embracing the social rei'ohiiiofi of the 1960s as the secondary topic of article 3. Approximately half of the respondents (39/82 = 48%) did not feel that there was a secondary topic in article 3 (see Table 20). For the 43 cases where participants reported a secondary topic, there was no consensus and only four responses matched the authors'. Seventy percent of these respondents (30/43) stated that they relied on the body of the text to help them in their task.

When the participants were asked how important some aspects of the subject were in relation to each other, the responses reflected the extent of agreement the participants had in identifying

Table 19
Article 3: Primary topic and factors helpful in determining it

Factor / Primary topic	Lay-out	Body	Layout and body	Sub-jec-tive	No resp.	Total
1	3	3	1	0	1	8
2	0	5	2	0	1	8
3	1	4	1	0	0	6
4	4	17	1	1	3	26
5	3	4	3	1	2	13
6	0	11	3	2	1	17
7	0	3	1	0	0	4
8	0	0	0	0	1	1
10	0	0	0	0	2	2
11	0	1	0	0	0	1
12	0	0	0	0	1	1
99	0	0	0	0	17	17
Total	11	48	12	4	29	104

Legend for "Primary topic":
1. Personal experiences in searching for social and cultural identification (authors' choice).
2. Effects of parenting by parents embracing the social revolution of the 1960s.
3. Generation gap.
4. Current self-perception of young people.
5. Life choices.
6. Twentysomething generation group.
7. A personal analysis.
8. Behaviour of parents.
10. Social or cultural change.
11. 1960s.
12. Women's issues.
99. No response.

the primary and secondary topics and, accordingly, the extent of variability in responses (see Table 21). Ninety-four percent (77/82) of the respondents considered that the primary topic was "more important," "a bit more important" or "as important" than the other aspects of the subject (i.e., for the most part the primary topic stood out as more important; therefore, making it easy to identify). It should be noted that more than 50% of the participants did not provide a response for the next two ratings of the relevant importance of secondary and tertiary topics. Seventy-four percent (37/50) of the respondents felt that the secondary topic was "as important" or "a bit less important" than the primary topic (i.e., the secondary topic was not considerably less important; therefore, making it difficult to identify). Seventy-one percent (35/49) of the respondents found that other aspects of the subject were

Table 20
Article 3: Secondary topic and factors helpful in determining it

Factor / Secondary topic	N/A	Lay-out	Body Body	Lay-out and body	Pers. knowl.	No resp.	To-tal
0	39	0	0	0	0	0	39
1	0	0	0	0	0	0	0
2	0	1	2	1	0	0	4
3	0	0	3	1	0	0	4
4	0	0	4	0	0	0	4
5	0	0	3	0	0	2	5
6	0	0	4	0	1	1	6
8	0	0	0	1	0	0	1
11	0	1	6	0	0	1	8
12	0	0	0	1	0	0	1
13	0	1	5	0	0	0	6
14	0	0	2	0	0	0	2
15	0	1	0	0	0	0	1
16	0	0	1	0	0	0	1
99	0	0	0	0	0	22	22
Total	39	4	30	4	1	26	104

Legend for "Secondary topic":
0, No secondary topic.
2, Effects of parenting by parents embracing the social revolution of the 1960s (authors' choice).
3, Generation gap.
4, Current self-perception of young people.
5, Life choices.
6, Twentysomething generation group.
8, Behaviour of parents.
11, 1960s.
12, Women's issues.
13, Lack of identity.
14, Cynicism.
15, Religious concerns.
16, Relationships.
99, No response.

Table 21
Article 3: Relative importance of different aspects of the subject

Topics / Rating	Primary vs. others	Secondary vs. primary	Secondary vs. others
Less important	0	5	4
A bit less important	5	12	7
As important	23	25	18
A bit more important	21	8	17
More important	33	0	3
No response	22	54	55

Table 22
Article 3: Ease in determining the general subject and reasons for the degree of ease

Reason / Ease	2	3	4	5	6	8	NR	Total
Very easy	0	0	2	2	1	0	1	6
Somewhat easy	0	1	4	0	5	1	2	13
Fair	1	0	0	0	4	5	8	18
Somewhat difficult	0	1	1	0	0	21	11	34
Very difficult	1	0	0	0	0	18	2	21
No response	0	0	0	0	0	0	12	12
Total	2	2	7	2	10	45	36	104

Legend for "Reasons":
2. Lack of familiarity with subject.
3. Familiarity with subject.
4. Layout.
5. Familiarity with subject and layout.
6. Clarity of narrative.
8. Lack of clarity of narrative.
NR. No response.

"a bit more important" and "as important" as the secondary topic (i.e., the relative importance of other topics apart from the primary one made it difficult to select one as secondary). The last two findings reveal why there was variability in the responses for a secondary topic for article 3.

Finally for article 3, the participants were asked how difficult it was to determine its subject, 60% (55/ 92) indicated that it was "somewhat difficult" to "very difficult" (see Table 22). The most frequent reason (39/ 55 == 71 %) these respondents gave for the relative difficulty in identifying the subject was the lack of clarity of the narrative. This finding emphasises the difficulty of determining the subject of a personal narrative and serves to demonstrate the potential in variability of responses for the general subject as well as primary and secondary topics of humanities text.

4. Discussion

The findings from novice indexers analysing the subject of three distinct articles reveal many differences as well as similarities in their performance and ease with the task. Article 1 was a social science text for which the majority of the respondents identified its general subject and primary topic, and found its general subject "somewhat easy" to "very easy" to determine. How ever, the responses regarding the secondary topic revealed less consensus. The performance of the novice participants is related to the discipline of the article, the factual nature of the text, the lack of complexity in the subject matter, the existence of various bibliographic apparatus, and the clarity of the text.

Article 2 was a science text for which the majority of the respondents identified its general subject and found it "somewhat easy" to "very easy" to determine. However, the responses regarding the primary and secondary topics indicated less consensus. The performance of the respondents is linked to the discipline of the article, the factual nature of the text, the complexity of the article's subject, the limited number of bibliographic apparatus, and the clarity of the text.

The third article was a humanities text which the majority of the respondents found "somewhat difficult" to "very difficult" to determine its subject and only a minority identified its general subject, primary and secondary topics. Many did not feel that article 3 had a secondary topic. The performance of the respondents is related to the discipline of the article, the subjective nature of the text, the relative complexity of the subject matter, the limited number of bibliographic apparatus, and the lack of clarity of the text.

Similar results were found for all three articles with regard to the factors which helped the participants determine the general subject and its various aspects. Here, the participants relied most heavily on the bibliographical apparatus and the layout. To determine the primary and secondary topics of the articles, the body of the texts was the most helpful.

The ease with which the participants determined the general subject was very high with the social science and science texts (articles 1 and 2) and very low with the humanities text (article 3). The relative ease participants experienced in determining the subject of the first article was attributed firstly to the layout of the article and secondly to the clarity of the narrative. The relative ease which participants had in analysing the subject of the second article was due to the clarity of the narrative first and the layout second. The participants felt that the relative difficulty they experienced in determining the subject of article 3 was due to the lack of clarity of the narrative.

This study posed two questions, one relating to the performance of the task of subject analysis and the second relating to the degree of ease in performing the task. Firstly, can novice indexers successfully analyse the subject of texts? The answer is overwhelmingly positive when determining the general subject of articles 1 and 2, a social science and a science text. However, only a minority of respondents (38%) were successful in determining the general subject of article 3, a humanities text. It should be noted that 28% reported a broader subject which indicates that the authors' definitive general subject had been taken into consideration.

Performance in determining the primary topic was mixed. The majority of the participants matched the authors' primary topic for article 1 but had poorer levels of performance with articles 2 and 3 - the former mainly because its subject was compound and the latter because the text was subjective and the little available bibliographic apparatus was uninformative, making it difficult to find a prominent topic.

In all three texts the performance in identifying the secondary topic was poor. In the first case, the primary topic was clearly prominent and the other aspects of the subject were considered to be of relative importance, making the task of selecting the secondary topic quite difficult. In the case of articles 2 and 3, the difficulties en countered in identifying the primary topic were compounded when determining the secondary topic.

The second research question dealt with the relative ease of performing subject analysis. An overwhelming majority of the respondents felt that it was quite easy to determine the general subject of articles 1 and 2. In contrast, the majority of the respondents experienced some difficulty in determining the general subject of article 3. This difference was attributed to the lack of clarity in the text of this personal narrative. The findings show that the overall performance in the task of subject analysis is successful when analysing the general subject of factual texts and the primary topic of texts with one central concept, and unsuccessful when the secondary topic is considered as important as other aspects of the subject, whether primary or tertiary. The level of ease in performing the task of subject analysis is high when bibliographic apparatus is available and its content is informative, and when the body of the text is clear. Inevitably, difficulties in conducting subject analysis are experienced in texts that lack clarity. Five factors were found that influenced the degree of performance and ease with subject analysis, they are:

1. Discipline of subject of texts —- articles representative of social science and science texts resulted in successful performance and ease in identifying the general concept and the primary topic when it was clearly central. The humanities article was associated with poor performance and difficulty with the task.
2. Factual LIS. subjective texts —- factual texts, representative of typical social science and science documents, encountered the same degree of performance and ease as described above. The subjective text, article 3, was associated with the same poor performance and difficulty as a humanities text.
3. (3) Complexity of subject of texts: simple or com-pound subjects —- compound subjects, as found in article 2, are associated with successful performance and relative ease when identifying the general subject but the contrary when prioritising the primary and secondary topics. Simple subjects are associated with successful performance and ease in identifying the general subject when the text is

factual and clear, e.g., article 1, but with poor performance and difficulties when determining the secondary topic because it is difficult to choose it from among other non-central aspects of the subject.

4. Presence of bibliographic apparatus —- the presence of more bibliographic apparatus (in particular title, sub-title, abstract, paragraph headings and initial paragraphs) with informative content resulted in successful performance and ease in analysing the general subject of texts having clarity.
5. Clarity of text —- clarity of the text was associated with a high degree of ease in determining the general subject of texts. The body of the text was helpful in the identification of primary and secondary topics while the layout was cited as more helpful in analysing the general subject.

The above findings have provided further insight into the process of subject analysis. Subject analysis at the macro-level is, for the most part, associated with success and ease in performance. At the micro-level, subject analysis is more problematic, its degree of success and ease are controlled by the influencing factors of each case. Although the cases were chosen to be representative of a variety of texts, it is necessary to clarify the external validity, intent and implications of the study.

5. Implications

The five factors which influence the process of subject analysis were presented separately but in fact are interrelated. From the descriptions above, one can detect the conspicuous connections. For example, from these findings, it appears that irrespective of subject discipline, with factual text, even if the subject is complex, bibliographical apparatus is a major factor in determining the general subject matter of text (articles 1 and 2). From the same results, it is also possible to observe that a humanities text which is less factual, which varies in its level of subject complexity, which has less bibliographic apparatus may be unclear and difficult to analyse (article 3). Such deductions must be approached with caution based merely on three texts which were chosen with some of these factors in mind.

However it is clear that combined with other factors, document-related factors do play a considerable role in aiding subject analysis. It is necessary to emphasise that the bibliographical apparatus considered in this study was typical of journal articles. The study has nothing to say about tables of contents, preface, etc. However, as it was prominent in helping pinpoint the general subject of the text, it would be interesting to discover from further research: (a) were they definitely the most important factor'? (b) did they just act as support to an initial reading of the text and (c) is there a discernible order of activity that could be determined?

The three documents were chosen to represent the disciplines of humanities, social sciences and sciences. The findings reinforce the notion that the two latter subjects are easier to analyse but no generalisations should be drawn from this. All articles were short, of a popular nature and required very little specialised knowledge.

There seems to be a serious problem when participants are required to isolate primary and secondary topics. This has to do with the process of prioritisation. A generalised statement may be easy or difficult to articulate, depending on several factors but when an attempt is made to push this analysis further, significant problems were encountered. In article 1 a clear, factual text. there was success in choosing the primary subject, yet in article 2, also a clear, factual text, there was less consensus. This appears to have been a question of balancing two topics of seemingly equal significance which resisted prioritisation. Inevitably, a secondary topic is difficult to isolate, as its identification depends on what has been chosen as primary. Clearly, more work needs to be done in this area, not because most texts need to have their subject content prioritised in this manner but because it has implications for issues of balance and specificity.

If subject analysis represents the first stages in the indexing process, then these findings highlight a series of issues which need to be addressed as part of the whole procedure. Given the possibility of expressing their concepts in natural language, participants still had some difficulties, especially when attempting to be specific (primary and secondary topics). Translating these concepts into an indexing language could lead to distortion if a badly articulated subject statement is fitted into an artificially controlled vocabulary and therefore could create information retrieval problems.

One aspect not studied in this research but worthy of future examination would be to match the terminology used by the participants against that of the texts examined, in the title, abstract, initial paragraphs, etc. An attempt to measure the extent to which words from the text were used could have some implications for research on extraction and other methods of automatic indexing.

The above findings and limitations of the study direct us to some areas of fruitful, basic research. In this age of information technology, the rush to automate processes should not be misinterpreted as the automation of a perfect indexing process. The continuing problems and questions of subject analysis, quality, effectiveness and efficiency of indexing, if not addressed, will be maintained and even compounded in an automated system, further hindering information retrieval and access.

References

[1] K.P. Jones, Towards a theory of indexing. Journal of Documentation 32(2) (1976) 118-125.

[2] E. Svenonius, Directions for research in indexing, classification, and cataloging. Library Resources & Technical Services 25(1)(1981)88-103.

[3] D.E. Kieras, A model of reader strategy for abstracting main ideas from simple technical prose. Text 2(1-3) (1982) 47-82.

[4] B.C. Vickery, Analysis of information. In: A. Kent and H. Lancour, eds. Encyclopedia of library and information science, vol. 1 (Dekker, New York, 1968) 355-384.

[5] C. Schwartz, Indexing behavior - survey and state of the art. In: B.M. Fry, ed. Information management in the 1980s: proceedings of the 40th ASIS Annual Meeting, vol. 14, Chicago, Illinois, September 26-October 1, 1977. Part 2 (on microfiche): Full papers (American Society for Information Science, White Plains, N.Y., 1977) Fiche 8, Frame D5-D14.

[6] W.J. Hutchins, The concept of "aboutness" in Subject indexing. Aslib Proceedings 30(5) (1978) 172-181.

[7] D.F. Swift, V. Winn and D. Bramer, "Aboutness" as a strategy for retrieval in the social sciences. Aslib Proceedings 30(5) (1978) 182-187.

[8] M.E. Maron, On indexing, retrieval and the meaning of about. In: S.K. Martin, ed. Information * politics: proceedings of the 39th ASIS Annual Meeting, vol. 13, San Francisco, California, October 4-9, 1976. Part 2 (on microfiche): Full papers (American Society for Information Science, Washington, D.C., 1966) Fiche 3, Frame E3.

[9] K.P. Jones, How do we index?: a report of some ASLIB Informatics Group activity, Journal of Documentation 39(1) (1983) 1-23.

[10] A. Rake, South Africa wants to make friends, New African (January 1991) 27.

[11] M.D. Lemonick, A restless venus unveiled, Time (October 8, 1990) 69.

[12] A. Berger, Too many options? A twentysomething woman ponders her future, Utne Reader (January/February 1991) 60-2. Excerpted from L.A. Weekly (March 30, 1990).